

Latent Factor Models for Density Estimation

Suprateek Kundu, Dept. of Biostatistics, UNC Chapel Hill, U.S.A. (skundu@live.unc.edu).

David B. Dunson, Dept. Statistical Science, Duke University, U.S.A. (dunson@stat.duke.edu).

Abstract: Although discrete mixture modeling has formed the backbone of the literature on Bayesian density estimation, there are some well known disadvantages. We propose an alternative class of priors based on random nonlinear functions of a uniform latent variable with an additive residual. The induced prior for the density is shown to have desirable properties including ease of centering on an initial guess for the density, large support, posterior consistency and straightforward computation via Gibbs sampling. Some advantages over discrete mixtures, such as Dirichlet process mixtures of Gaussian kernels, are discussed and illustrated via simulations and an epidemiology application.

Keywords: Nonparametric Bayes; Kernel estimation; Density regression; Gaussian process; Latent variable model; Dirichlet process; Posterior consistency; Latent factor regression.

1. INTRODUCTION

Nonparametric kernel mixture models are increasingly popular in density estimation and high dimensional data modeling. Kernel mixture models have the form:

$$f(y; G) = \int \mathcal{K}(y; \theta) G(d\theta), \quad (1)$$

where $G(\cdot)$ is a mixing distribution and $\mathcal{K}(\cdot)$ is a probability kernel. The majority of the nonparametric Bayesian development in this area relies on Dirichlet process (DP) priors (Ferguson, 1973; 1974) for G . These models have been generalized to density regression by defining dependence on the covariates x in various ways. Müller, Elkanli and West (1996) used a DP mixture of multivariate normals to jointly model the density of the response and predictors to induce a prior on $f(y|x)$. In order to let the parameters of the DP vary over the predictor space \mathcal{X} , MacEachern (1999) defined dependent Dirichlet processes (DDP) by assigning stochastic processes on the components in Sethuraman's (1994) DP representation: $G_x = \sum_{i=1}^{\infty} p_i(x) \delta_{\theta_i(x)}$. De Iorio et al. (2004) proposed a fixed- p DDP, while Griffin and Steel (2006) allowed the weights to depend on predictors. Dunson, Pillai and Park (2007) instead used predictor-dependent convex combinations of DP components.

There is also a rich literature on using mixture priors in hierarchical latent variable models. Bush and MacEachern (1996) and Kleinman and Ibrahim (1998) proposed DP mixtures on the distributions of random effects. Fokoue and Titterington (2003) and Fokoue (2005) proposed mixtures of factor analyzers (MFA) corresponding to a finite mixture of multivariate normal kernels with a factor-analytic decomposition of the component-specific covariances. Dunson (2006) used dynamic mixtures of DPs to allow a latent variable distribution to change nonparametrically across groups. More recently, Chen et al. (2009) and

Carvalho et al. (2008) proposed nonparametric Bayes MFA allowing an uncertain number of factors. Lee, Lu, and Song (2008) placed a truncated DP on the distribution of the latent variables within a structural equation model (SEM), while Yang and Dunson (2010) proposed a centering approach to ensure identifiability of the latent factor distributions.

The above approaches have relied on discrete mixture models, which have a number of well known complications motivating alternative methods for modeling unknown densities, such as Polya trees (Mauldin et al, 1992; Lavine, 1992, 1994) and logistic Gaussian processes (LGP) (Lenk 1988, 1991; Tokdar 2007). Polya trees have appealing properties in terms of denseness, conjugacy and posterior consistency but have disadvantages in terms of favoring overly spiky densities. LGP has sound theoretical properties and smoothness of the densities can be controlled through the covariance kernel in the GP. However, posterior computation is a major hurdle. Recently, Jara and Hanson (2010) proposed dependent tail-free processes where they modeled the tail-free probabilities with LGP dependent on covariates. Their approach is shown to approximate the Polya tree marginally at each predictor value. An alternative was suggested by Tokdar, Zhu and Ghosh (2010) relying on LGP for density regression with dimensionality reduction.

In this article, we focus on a new approach for nonparametric density estimation and regression that induces a prior on the unknown density through placing a flexible prior on a nonlinear regression function θ in a latent factor model. The proposed class of models is related to Gaussian process latent variable models (GP-LVM) proposed in the machine learning literature (Lawrence, 2005; Silva and Gramacy, 2010), but our modeling details are different and the focus of this literature has been on nonlinear dimensionality reduction with no consideration of density estimation and associated properties. By using GP priors for θ ,

we obtain substantial control over the smoothness of the induced densities in a very different manner than that achieved by LGP-based models. Unlike LGP-based models, the proposed model has desirable conjugacy properties facilitating posterior computation. In addition, the method has appealing theoretical properties in terms of large support and posterior consistency.

Relative to some density estimation priors, the proposed latent factor approach is quite easy to generalize to more challenging settings involving multivariate densities, conditional density estimation, hierarchical modeling and other complexities. Although our primary focus in this article is to introduce the formulation, providing an intuition for how the model works, basic properties and computation, we also give a flavor of generalizations through a simple conditional density estimation example. In particular, we consider a model that induces a prior on the conditional density $f(y|x)$ through joint modeling of the response and predictors through separate nonparametric latent factor models containing the same latent variables. This formulation is completely flexible in the marginal densities, while making strong restrictions on the dependence to address the curse of dimensionality in a related manner to a copula model. An attractive feature of our model is that it naturally allows for incorporation of prior information on the marginal densities of response and predictors through the mean function of the GP. The utility of incorporating such prior information is clear in a reproductive epidemiology application we consider.

2. DENSITY ESTIMATION

2.1. Model Specification

Initially suppose y_i are iid draws from an unknown density $f \in \mathcal{F}$, where \mathcal{F} is the set of

densities on \mathfrak{R} with respect to Lesbesgue measure. We propose to induce a prior $f \sim \Pi$ through

$$\begin{aligned} y_i &= \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \Gamma_\sigma, \\ \mu &\sim \Pi^*, \quad \sigma \sim \nu, \quad x_i \sim \text{Uniform}(0, 1), \end{aligned} \tag{2}$$

where $\mu \in \Theta$ is an unknown $[0, 1] \rightarrow \mathfrak{R}$ function, x_i is a uniformly distributed latent variable, and the error distribution Γ_σ is centered at 0 and has scale parameter σ . Hence, in the special case in which $\mu(x) = \mu$, so that the regression function is a constant, and Γ_σ is normal, we have $f(y; \mu, \sigma^2) = N(y; \mu, \sigma^2)$ so we obtain a normal density. The density of y conditionally on the unknown regression function μ and σ is obtained on marginalizing out the latent variable as

$$f(y; \mu, \sigma) = f_{\mu, \sigma}(y) = \int_0^1 \Gamma_\sigma(y - \mu(x)) dx. \tag{3}$$

To complete the specification, we let $\mu \sim \Pi^*$, $\sigma \sim \nu$ and obtain the marginal density

$$f(y) = \int_0^\infty \int_\Theta \int_0^1 \Gamma_\sigma(y - \mu(x)) dx \Pi^*(d\mu) \nu(d\sigma). \tag{4}$$

Hence, a prior $f \sim \Pi$ is induced through assigning independent priors to μ and σ in expression (3). When the prior on μ is a Gaussian process and the error distribution is $N(0, \sigma^2)$ (denoted as $\Gamma_\sigma = \phi_\sigma$), we refer to $f_{\mu, \sigma}$ as a Gaussian process transfer (GPT) model and the induced prior $f \sim \Pi$ as a GPT prior.

The GPT prior does not have the kernel mixture form (1). There will be no clustering of subjects or label switching issues. Instead, the prior $f \sim \Pi$ is induced through adding a Gaussian residual to a Gaussian process regression model in a uniform latent variable. This is a simple structure aiding computation and interpretability. One can control the smoothness

of the density through the covariance in the GP prior for the regression function μ and the size of the scale parameter σ . In limiting cases, one can obtain realizations of μ concentrated close to a flat line, leading to a normal density as a special case. In addition, by making σ small and choosing the GP covariance to generate a very bumpy μ , one can obtain arbitrarily bumpy densities. In practice, by choosing hyperpriors for key covariance parameters, we obtain a data adaptive approach that often outperforms discrete kernel mixtures. The performance of discrete kernel mixtures relies on the ability to accurately approximate the density with few components, and DP mixtures tend to heavily favor a small number of dominate kernels. This tendency can sometimes lead to relatively poor estimation, as illustrated in section 6.

2.2. Prior Specification

Prior elicitation is an important aspect of Bayesian modeling, with the prior playing a particularly important role in Bayesian nonparametric models involving infinitely many parameters. Most of the Bayesian nonparametrics literature relies on default priors, which do not reflect available prior knowledge in a particular application area, but are chosen to lead to good performance in terms of posterior behavior in a wide variety of applications. However, as in parametric models, well chosen informative priors that utilize information, such as historical data on the variables under study, can substantially improve the performance in small to moderate samples. In DP mixtures, such prior information is typically incorporated through choice of hyperparameters in the base measure, while maintaining conjugacy for ease in computation. For example, in Gaussian kernel mixtures, a normal-inverse gamma base measure would be chosen having parameters representing prior knowledge. This is appropriate when prior knowledge implies that the density follows a t distribution, but when one has

prior information that the density follows a more complex form (as in our premature delivery application) then elicitation is substantially more difficult. Obtaining a base measure that leads to a particular elicited density is a deconvolution problem, which can be difficult to solve for non-atomic base measures. In addition, posterior computation under the resulting complex and non-conjugate base measure may be challenging. An advantage of the GPT is that the prior for the density can be centered on an arbitrary choice easily through the prior mean in the GP prior for μ .

To elaborate, Theorem 3 (section 3) ensures that $f_{\mu,\sigma} \approx f_{\mu_*,\sigma}$, when $\mu \approx \mu_* = F_*^{-1}$ and $\sigma \approx 0$. In terms of application, this translates to incorporating a prior guess $f_{\mu_*,\sigma}$ for the density through the corresponding mean function $\mu_* = F_*^{-1}$ of the GP, and letting the prior for σ to have mode near 0. Such a mean function can be constructed by obtaining frequentist kernel estimates of the concerned density using some external data, and then converting it into an inverse cdf on a grid of points in $[0,1]$ (using a linear approximation). Thus, the characteristics of the entire density is captured through F_*^{-1} (mean function of the GP) and we let the data influence the deviation of the posterior from the prior guess.

These ideas are demonstrated in Figure 1, where we use some earlier data on gestational age at delivery to construct prior densities. We choose a $\text{Ga}(25,1)$ prior for the residual precision, and different sets of hyper-parameters for the covariance kernel of the GP. The frequentist kernel estimates were obtained by the bandwidth selection method of Sheather and Jones (1991), using a Gaussian kernel ('kernel' function in R). It is evident that the smoothness as well as the degree of deviation of the prior from the frequentist estimate can be controlled through the hyper-parameters in the covariance kernel of the GP.

3. Theoretical Properties

To further justify the proposed prior, we show large support and posterior consistency properties. Large support is an important property in that it ensures that our prior can generate densities that are arbitrarily close to any true density f_0 in a large class, a defining property for a nonparametric Bayesian procedure and a necessary condition to allow the posterior to concentrate in small neighborhoods of the truth. Instead of focusing narrowly on GPT priors, we provide broad theoretical results for priors in the general class of expression (2).

Before proceeding, it is necessary to define some notation and concepts. We denote the Kullback-Leibler (KL) divergence of $f_{\mu,\sigma}$ from f_0 as $KL(f_{\mu,\sigma}, f_0)$ and an ϵ -sized KL neighborhood around f_0 as $KL_\epsilon(f_0)$. The sup-norm distance is denoted by $\|\cdot\|_\infty$. Our development will rely on the fact that any density f_0 can be calculated from a “true function” $\mu_0 = F_0^{-1}$, with F_0 denoting the cumulative distribution function. To generate $y_i \sim f_0$, one can equivalently draw $x_i \sim \text{Uniform}(0, 1)$ and let $y_i = \mu_0(x_i)$. This corresponds to the limiting case as $\sigma \rightarrow 0$ in model (2) with $\mu = \mu_0$. Hence, we assume a weak regularity condition that the true density can be represented as the limiting case

$$f_0(y) = \lim_{\sigma \rightarrow 0} \int_0^1 \Gamma_\sigma(y - \mu_0(x)) dx, \quad (5)$$

assuming Γ_σ is chosen so that such a limit exists. The above condition is quite reasonable, only excluding densities for which convergence in distribution does not imply convergence of the corresponding density functions. As additional reasonable regularity conditions, we assume that f_0 is strictly positive and finite, $\sup_x |\mu_0(x)| < \infty$ and $0 < \Gamma_\sigma(u) < \infty$ for finite u . A uniformly bounded μ_0 along with the condition on Γ_σ implies that for μ belonging to a ball of finite radius around μ_0 , $f_{\mu,\sigma}$ is strictly positive and finite for all $\sigma \in \mathbb{R}^+$, which ensures a finite KL divergence for a suitable subset of μ values.

Theorem 1. *Let $\sup_x |\mu_0(x)| < \infty$, Γ_σ be normal, Laplace or Cauchy with scale parameter σ and f_0 be the corresponding density in \mathcal{F} defined as in equation (5). If μ_0 is in the sup-norm support of Π^* and $\nu\{\sigma : \sigma \in (0, \eta)\} > 0$ for all $\eta > 0$, then $\Pi(KL_\epsilon(f_0)) > 0$ for all $\epsilon > 0$.*

Theorem 1 allows us to verify that the induced prior on the density f assigns positive probability to KL neighborhoods of any strictly positive and finite true density f_0 . From Schwartz (1965), if the true density f_0 is in the KL support of the prior for f , the posterior distribution for f will concentrate asymptotically in arbitrarily small weak neighborhoods of f_0 . Theorem 1 requires the prior $\mu \sim \Pi^*$ to place positive probability in sup-norm neighborhoods of the inverse cdf F_0^{-1} . Although one can verify this condition for certain choices of Π^* , such as appropriately chosen Gaussian process priors, it is nonetheless somewhat stringent. We show in Theorem 2 that this condition can be relaxed to only require that the prior $\mu \sim \Pi^*$ assigns positive probability to L-1 neighborhoods of any element μ_0 of Θ . It is well known that positive sup-norm support automatically guarantees positive L-1 support but the converse is not true.

Theorem 2. *Let $\sup_x |\mu_0(x)| < \infty$, $\Gamma_\sigma = \phi_\sigma$ and f_0 be the corresponding density in \mathcal{F} defined in equation (5). If μ_0 is in the L-1 support of Π^* and $\nu\{\sigma : \sigma \in (0, \eta)\} > 0$ for all $\eta > 0$, then $\Pi(KL_\epsilon(f_0)) > 0$ for all $\epsilon > 0$.*

As the prior $f \sim \Pi$ is specified indirectly through priors $\mu \sim \Pi^*$ and $\sigma \sim \nu$, it is desirable for elicitation purposes to verify that, for sufficiently small σ , $\mu \approx \mu_0$ implies that $f_{\mu, \sigma} \approx f_0$. Theorem 3 provides such a verification assuming Gaussian errors. This implies one can potentially center the prior for the density f on an initial parametric guess \tilde{f} by centering $\mu \sim \Pi^*$ on the inverse cdf \tilde{F}^{-1} while choosing the prior for σ to have mode near zero. The

data will then inform about the degree to which μ deviates from \tilde{F}^{-1} and σ deviates from 0.

Theorem 3. *For $\mu_0 \in \Theta$ and $\Gamma_\sigma = \phi_\sigma$, let f_0 be the density resulting from equation (5). Then for $\mu \in N_{\epsilon_1}(\mu_0)$, with $N_{\epsilon_1}(\mu_0)$ an ϵ_1 -sized L-1 neighborhood around μ_0 , and $\sigma \in (\epsilon_2, \epsilon_2^*)$, we have $f_{\mu,\sigma} \in N_{\frac{\epsilon_1}{\epsilon_2}}(f_0)$ for arbitrarily small $\epsilon_1, \epsilon_2, \epsilon_2^*$ such that $0 < \epsilon_1 < \epsilon_2 < \epsilon_2^*$.*

Although Theorems 1-2 lead to weak posterior consistency, small weak neighborhoods around f_0 are topologically too large and may include densities that are quite different from f_0 in shape and other characteristics. Hence, it is appealing to establish a strong posterior consistency result in which the posterior probability allocated to arbitrarily small L-1 neighborhoods of f_0 increases towards one exponentially fast with increasing sample size. Focusing on the GPT prior described above, we show in Theorem 4 that strong posterior consistency holds under some conditions on the prior. Notably, for an appropriately chosen GPT prior, we obtain L-1 posterior consistency for all strictly positive and finite true densities $f_0 \in \mathcal{F}$ with the weak regularity condition (5).

Theorem 4. *Suppose $\mu \sim GP(m, c)$ and define f_0 as in equation (5) where $\Gamma_\sigma = \phi_\sigma$. Let $U = \{f_{\mu,\sigma} : \int |f_{\mu,\sigma} - f_0| dy < \epsilon, \mu \in \Theta, \sigma \in (0, \infty)\}$. Suppose the mean function $m(\cdot)$ is continuously differentiable, and the covariance function $c(\cdot, \cdot)$ has continuous fourth derivatives. Further $0 < \nu(\sigma : \sigma \in (0, \eta_1)) < \eta_2$ for any arbitrarily small η_1 , with η_2 decreasing with η_1 . Then, f_0 is in the KL support of Π implies that the posterior is strongly consistent at f_0 .*

The above assumptions on the GP can be verified for many popular covariance functions, both stationary and nonstationary. Some such examples can be found in Choi et. al. (2004).

4. SINGLE FACTOR DENSITY REGRESSION

As a simple and parsimonious single factor model that generalizes the model of Section 2 to

include predictors $z_i = (z_{i1}, \dots, z_{ip})'$ of a response y_i , we let

$$\begin{aligned}
y_i &= \mu^Y(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_Y^2), \\
z_{ik} &= \mu^{Z_k}(x_i) + \epsilon_{ik}^*, \quad \epsilon_{ik}^* \sim N(0, \sigma_{Z_k}^2), k = 1, 2, \dots, p, \\
x_i &\sim \text{Uniform}(0, 1), \\
\mu^Y &\sim \Pi^Y, \quad \mu^{Z_k} \sim \Pi^{Z_k}, k = 1, 2, \dots, p, \\
\sigma^Y &\sim \nu, \quad \sigma_{Z_k} \sim \nu,
\end{aligned} \tag{6}$$

where $\mu^Y, \mu^{Z_k} \in \Theta$ are unknown $[0, 1] \rightarrow \Re$ functions, ϵ 's are independent errors and x_i is the latent variable. For simplicity, we assume the same prior ν on the precision of the measurement errors in each component model, though this assumption is trivial to relax. Expression (6) is a multivariate generalization of the univariate density estimation model (2). Marginally each of the variables is assigned exactly the prior in (2) and to allow dependence we incorporate the same latent factor x_i in each of the models.

Our goal in defining a joint model is to induce a flexible but parsimonious model for the conditional density of y_i given the predictors \mathbf{z}_i . In estimating conditional densities for multiple predictors, one encounters a daunting dimensionality problem in that one is attempting to estimate a density nonparametrically while allowing arbitrary changes in this density across a multivariate predictor space. Clearly, as p increases even for large samples there will be many regions of the predictor space that have sparse observations. As a compromise between flexibility and parsimony in addressing the curse of dimensionality, we propose to use a single factor model in which the marginals for each variable are fully flexible but restrictions come in through assuming dependence on a single x_i . Extensions to the multiple factor case are straightforward.

5. POSTERIOR COMPUTATION

For simplicity, we focus on the single predictor density regression case when outlining an MCMC algorithm for posterior computation. Let $\mathbf{Y}_{n \times 1}$ and $\mathbf{Z}_{N \times 1}$ denote the vector of observations and covariates, respectively. We are interested in prediction of y_{n+1}, \dots, y_N based on z_{n+1}, \dots, z_N . Let μ_Y^n ($n \times 1$) and μ_Z^N ($N \times 1$) denote the realizations of the GP μ^Y and μ^Z at the latent variable values $\mathbf{x} = (x_1, \dots, x_n, x_{n+1}, x_N)'$. From the GP prior, we have $\mu_Y^n \sim N_n(m_Y^n, \mathbf{K}_Y^n)$ and $\mu_Z^N \sim N_N(m_Z^N, \mathbf{K}_Z^N)$. The covariance kernels are squared exponential with $\mathbf{K}_Y(x, x') = \frac{1}{\phi_Y} \exp \left\{ -C_Y(x - x')^2 \right\}$ and $\mathbf{K}_Z(x, x') = \frac{1}{\phi_Z} \exp \left\{ -C_Z(x - x')^2 \right\}$. We specify conjugate gamma priors: $\sigma_Y^{-2} \sim Ga(a_\sigma, b_\sigma)$, $\sigma_Z^{-2} \sim Ga(aa_\sigma, bb_\sigma)$, $\phi_Y \sim Ga(a_\phi, b_\phi)$ and $\phi_Z \sim Ga(aa_\phi, bb_\phi)$. For updating the latent variables \mathbf{x} , we adopt the griddy Gibbs approach using a set of evenly distributed grid points $g_1^*, g_2^*, \dots, g_G^* \in (0, 1)$. Let \mathbf{D}_Y and \mathbf{D}_Z be diagonal matrices having σ_Y^2 and σ_Z^2 as their diagonal elements respectively. Let $\mu_Y^n(-i)$ include all elements of μ_Y^n except $\mu^Y(x_i)$, and similarly for $\mu_Z^N(-i)$. The Gibbs sampling algorithm alternates between the following steps.

Step1: Update σ_Y^2 and σ_Z^2 using $\pi(\sigma_Y^{-2} | -) \sim Ga(a_\sigma + n/2, b_\sigma + \frac{1}{2} \sum_{i=1}^n (y_i - \mu^Y(x_i))^2)$ and $\pi(\sigma_Z^{-2} | -) \sim Ga(aa_\sigma + N/2, bb_\sigma + \frac{1}{2} \sum_{i=1}^N (z_i - \mu^Z(x_i))^2)$ respectively.

Step2: To sample the latent variables, choose $x_i = g_k^*$ with probability p_{ik} , where

$$\begin{aligned} p_{ik} = P[x_i = g_k^* | -] &= \frac{p_{ik}^Y p_{ik}^Z N(\mu^Y(x_i = g_k^*) | \mu_Y^n(-i)) N(\mu^Z(x_i = g_k^*) | \mu_Z^N(-i))}{\sum_{l=1}^G p_{il}^Y p_{il}^Z N(\mu^Y(x_i = g_l^*) | \mu_Y^n(-i)) N(\mu^Z(x_i = g_l^*) | \mu_Z^N(-i))}, \text{ if } i \leq n \\ &= \frac{p_{ik}^Z N(\mu^Z(x_i = g_k^*) | \mu_Z^N(-i))}{\sum_{l=1}^G p_{il}^Z N(\mu^Z(x_i = g_l^*) | \mu_Z^N(-i))}, \text{ if } n < i \leq N, \end{aligned}$$

where $p_{ik}^Y = N(y_i; \mu^Y(x_i = g_k^*), \sigma_Y^2)$, $p_{ik}^Z = N(z_i; \mu^Z(x_i = g_k^*), \sigma_Z^2)$ and $k=1, 2, \dots, G$.

Step3: Update μ_Y^n and μ_Z^N using $\pi(\mu_Y^n | -) = N_n \left((\mathbf{D}_Y^{-1} + (\mathbf{K}_Y^n)^{-1})^{-1} (\mathbf{D}_Y^{-1} \mathbf{Y} + (\mathbf{K}_Y^n)^{-1} m_Y^n), (\mathbf{D}_Y^{-1} + (\mathbf{K}_Y^n)^{-1})^{-1} \right)$ and $\pi(\mu_Z^N | -) = N_N \left((\mathbf{D}_Z^{-1} + (\mathbf{K}_Z^N)^{-1})^{-1} (\mathbf{D}_Z^{-1} \mathbf{Z} + (\mathbf{K}_Z^N)^{-1} m_Z^N), (\mathbf{D}_Z^{-1} + (\mathbf{K}_Z^N)^{-1})^{-1} \right)$

respectively.

Step4: Update $\mu_Y^{*G} = \{\mu^Y(g_1^*), \dots, \mu^Y(g_G^*)\}$ and $\mu_Z^{*G} = \{\mu^Z(g_1^*), \dots, \mu^Z(g_G^*)\}$ using the conditional normal distributions $N(\mu_Y^{*G}|\mu_Y^n)$ and $N(\mu_Z^{*G}|\mu_Z^N)$ respectively.

Step5: Update ϕ_Y and ϕ_Z using $\pi(\phi_Y|-) \sim \text{Ga}(a_\phi + \frac{n}{2}, b_\phi + \frac{1}{2}(\mu_Y^n - m_Y^n)'(\mathbf{K}_Y^n)^{-1}(\mu_Y^n - m_Y^n))$ and $\pi(\phi_Z|-) \sim \text{Ga}(aa_\phi + \frac{N}{2}, bb_\phi + \frac{1}{2}(\mu_Z^N - m_Z^N)'(\mathbf{K}_Z^N)^{-1}(\mu_Z^N - m_Z^N))$ respectively.

Step6: Update C_Y and C_Z using Metropolis random walk for $\log(C_Y)$ and $\log(C_Z)$.

For prediction of y_k based on z_k , $k = n+1, \dots, N$, we use $\pi(y_k|-) = N(y_k; \mu^Y(x_k), \sigma_Y^2)$, while the conditional density estimate is calculated as $\hat{f}(y|z) = \frac{\frac{1}{G} \sum_{k=1}^G \phi_{\sigma_Y}(y - \mu^Y(g_k^*)) \phi_{\sigma_Z}(z - \mu^Z(g_k^*))}{\frac{1}{G} \sum_{k=1}^G \phi_{\sigma_Z}(z - \mu^Z(g_k^*))}$.

6. SIMULATION STUDY

To assess the performance of the GPT approach in density estimation as well as density regression, we conducted several simulation studies. We chose the mean function for the GP as $m(x) = 2\sin(x) + \cos(x)$ and utilized the squared exponential covariance kernel. For computational purposes, we worked with the standardized data and then transformed it back in the final step. The hyperparameters for the gamma priors were chosen to be one throughout. Although we used 75 grid points for the griddy Gibbs approach, the number of points could be as low as 60. The number of iterations used was 10000 with a burn in of 1000. The convergence for the main quantities such as μ was rapid with good mixing. All results are reported over 5 replicates.

6.1. Univariate Density Estimation

To see how well the GPT does in practice for density estimation, we looked at a variety of scenarios, where the truth was generated from the densities considered in Marron and Wand (1992), which are essentially finite mixtures of Gaussians. We present the results from four of those cases which we thought to be interesting deviations from normality and could be

potentially encountered in applications. These are the 2nd, 6th, 8th and 9th Marron-Wand densities. The sample size used was 100. For comparison, we looked at DP mixture of Gaussians (Escobar and West, 1995), mixtures of Polya trees (Hanson, 2006) and frequentist kernel estimates using a Gaussian kernel (and the bandwidth selection method of Sheather and Jones, 1991). More specifically, for both DP mixtures and mixtures of Polya trees, we used the DP package in R and the standard hyperparameter values therein. We used algorithm 8 of Neal (2000) with $m=1$ for DP mixtures of Gaussians. For frequentist kernel, we used the function “density” in R with Gaussian kernel. Overall, we found that varying the hyperparameter values within a reasonable range does not significantly alter the density estimation results for a sample size of 100, for any of the competitors. Table 1 presents the L-1 distance between true and estimated densities while Figure 2 depicts the density plots.

Table 1: Marron-Wand Curves: L-1 distance between true and estimated densities

Method	L-1 Distance			
	MW 2	MW 6	MW 8	MW 9
GPT	0.031	0.035	0.031	0.028
DPM	0.035	0.036	0.03	0.038
Polya tree mixture	0.065	0.036	0.045	0.042
Frequentist Kernel	0.145	0.031	0.033	0.028

From table 1, we see that even when the truth is generated from a finite mixture of Gaussians, the GPT tends to do better or at least as well as the DP mixture of Gaussians. Mixtures of Polya trees have somewhat worse performance and result in overly spiky looking estimates.

6.2. Single Factor Density Regression

For density regression, we generated a univariate response by allowing the conditional mean as well as the residual error distribution to vary with the covariate. We compared the out of sample predictive performance of GPT with other competitors such as DP mixture of bivariate normals (Müller, Erkanli and West, 1996), Bayesian additive regression trees (BART) (Chipman, George and McCulloch, 2010), GP mean regression (O’Hagan and Kingman,

1978) and treed GP (Gramacy and Lee, 2008), based on standard packages in R. We used the DP package for DP mixtures of Gaussians and the Bayestree package for the other three methods, and the hyperparameter values therein. The density regression results did not change significantly on varying the hyperparameter values within a reasonable range, for all the competitors. We used the following scheme for simulations:

$$Z \sim F_Z, \quad y_i = \lambda \exp\left(-\frac{e^{z_i}}{1 + e^{z_i}}\right) + \frac{e^{z_i}}{1 + e^{z_i}} \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

where F_Z is the distribution of the predictors which was chosen to be a trimodal density (9th Marron-Wand curve). We chose $\lambda = 3$ and split the total sample size of 100 into training set of 50 and test set of 50. The above data generating model allows the shape of the conditional density to change with predictors, hence making prediction non-trivial. Table 2 shows the performance of the GPT along with a few competitors. We computed the mean square error (MSE), 95% coverage for the mean (COV), as well as the L-1 distance between true and estimated densities at 25th, 50th and 75th percentiles of the predictor distribution.

Table 2: Out of Sample MSE and L-1 distance between true and estimated densities

Method	MSE	COV(%)	L-1 Distance		
			25th	50th	75th
GPT	1.26	94	0.08	0.04	0.06
DPM	1.53	42	0.03	0.04	0.06
BART	1.59	46	0.13	0.026	0.10
GP reg	1.52	76	0.14	0.03	0.10
treed GP	1.6	72	0.07	0.09	0.04

The results in table 2 are consistent with our experience in simulations- when the predictor distribution is multimodal and the shape of the conditional density is allowed to change with predictors, then the GPT tends to do as well or better than DP mixture of Gaussians. For the above study, the average number of components in the conditional distribution obtained from DP mixtures was around 15 which is quite high for a sample size of 50. As illustrated in

table 2, BART, treed GP and the GP mean regression methods are primarily mean regression methods and so cannot possibly do well in terms of characterizing the entire conditional of response given predictors. They might perhaps estimate the mean surface reasonably well, but eventually fail in capturing multimodality or tail behavior, the latter often being an important focus of inferences.

7. EPIDEMIOLOGY APPLICATION

7.1 Study Background

DDT is a cheap and popular alternative for reducing the transmission of malaria, but has been shown to have negative effects on public health. In order to study the association between the DDT metabolite DDE and preterm delivery, Longnecker et al. (2001) measured DDE in mother’s serum in the third trimester of pregnancy and also recorded gestational age at delivery (GAD) as well as age. They did logistic regression with response as dichotomized GAD (preterm or normal depending on a cut-off of 37 weeks of completed gestation) and explanatory variables as categorized DDE based on empirical quantiles. Their results showed a significant dose-response relationship which had important public health implications. Dunson et al. (2008) analyzed the data using kernel stick-breaking processes, and showed an increasing bump in the left tail of the GAD density with increasing DDE.

7.2 Analysis and Results

We used the GPT to analyze the dose response relationship in a subset of 182 women of advanced maternal age (≥ 35 yrs) in the above dataset. We examined the conditional distribution of GAD at 10th, 60th, 90th and 99th percentile of DDE. Further, we looked at the dose response relationship between preterm birth and DDE, by examining the left tail

of GAD over varying doses of DDE. We used normalized data for analysis and converted it back in the final step. Using the prior specification approach of section 2.2, we were able to incorporate prior information on the marginal density of GAD (using an external data) through the mean function of the GP. Note that prior on σ^{-2} for GAD was chosen as $\text{Ga}(25,1)$. Given the limited sample size and the complexity of the data we are trying to model, we adjusted other hyperparameter settings to reflect our prior belief about the data. The starting value for the length-scale parameter in the covariance kernel in the Metropolis random walk was chosen to be 25, so as to have smooth Gaussian process prior. Instead of working with DDE, we used $\log(\text{DDE})$ which resembled a Gaussian distribution, with a 0 mean function for the predictor component and $\text{Ga}(1,1)$ prior for the corresponding residual precision.

Figure 3 shows the conditional distribution curves for GPT along with 90% credible intervals. Although we focused on a small subsample of 182 women of advanced maternal age, the GPT results for the conditional density are remarkably similar to the ones reported in Dunson et al. (2008), which suggests that there is no systematic difference for women of advanced maternal age. The conditional densities show an increasing bump in the left tail with increasing DDE, suggesting increased risk of preterm birth at higher doses. This is further supported by dose-response curves for $P(\text{GAD} < T)$ in Figure 5, with different choices for cut-off T . Although the dose-response curve is mostly flat for $T=33$ weeks, the relationship becomes more significant as cut-off increases, with the dose-response tapering off at $T=40$ weeks. This suggests that increased risk of preterm birth at higher DDE dosage is attributable to premature deliveries between 33 and 37 weeks. Trace plots of $f(y|z)$ for different DDE percentiles (not shown) exhibit excellent rates of convergence and mixing.

For comparison, Figure 4 shows the density estimates from the DP mixture of Gaussians which has a tendency to overly favor multimodal densities, which is as expected given our simulation study results. These results were obtained using DP package in R (and the data driven hyperparameter values therein), which utilizes algorithm 8 of Neal (2000) with $m=1$.

8. Discussion

In this paper, we propose a latent factor model for density estimation. This novel method provides us with a flexible non-discrete mixture alternative to be used in a variety of situations including density estimation, density regression, hierarchical latent variable models and even mixed models. We provide theoretical justifications for GPT and demonstrate its usefulness as a building block for more complex models involving covariates. Building on our work, Pati, Bhattacharya and Dunson (to be submitted) recently showed minimax optimal rates of posterior contraction for Bayesian density estimation from non-linear latent variable models, also obtaining initial results on contraction rates in conditional density estimation. The close relationship between non-linear latent variable models for densities and non-linear mean regression models facilitates not only posterior computation but also derivations of theoretical properties, such as contraction rates, which have proven difficult to study for discrete mixtures beyond simple settings.

9. Acknowledgments

The authors thank Debdeep Pati and Anirban Bhattacharya for their helpful comments. This work was supported by Award Number R01ES017240 from the National Institute of Environmental Health Sciences. The content is solely the responsibility of the authors and

does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health.

APPENDIX: PROOF OF RESULTS

Proof of Theorem 1:

That $f_{\mu,\sigma}$ exists is seen from equation (3). The limiting case as $\sigma \rightarrow 0$ can be explained as the case when Y has the distribution function $F = \mu^{-1}$ and marginalizing out the latent variable $x \in [0, 1]$. Thus $\lim_{\sigma \rightarrow 0} \int_0^1 \Gamma_\sigma(y - \mu(x)) dx$ exists. Under regularity conditions, we can write the KL divergence between $f_{\mu,\sigma}$ and f_0 as

$$KL(f_{\mu,\sigma}, f_0) = \int f_0 \log \frac{f_0}{f_{\mu,\sigma}} = \int f_0(y) \left[\frac{\lim_{\sigma \rightarrow 0} \int_0^1 \Gamma_\sigma(y - \mu_0(x)) dx}{\int_0^1 \Gamma_\sigma(y - \mu(x)) dx} \right] dy. \quad (7)$$

Note, $\Gamma_\sigma(y - \mu(x)) \geq \frac{\Gamma_\sigma(y - \mu_0(x))}{h_\sigma(y, \mu(x) - \mu_0(x))}$, where $h_\sigma(y, \mu - \mu_0) = e^{\frac{1}{2\sigma^2}(\mu - \mu_0)^2 - \frac{1}{\sigma^2}(y - \mu_0)(\mu - \mu_0)}$ for normal error, while for Laplace error, $h_\sigma(y, \mu - \mu_0) = e^{\frac{1}{\sigma}(|\mu_0 - \mu|)}$. Further $\sup_x |\log(h_\sigma(y, \mu(x) - \mu_0(x)))| \rightarrow 0$ for all $y \in \mathbb{R}$ as $\|\mu - \mu_0\|_\infty, \sigma^2$ go to 0 with $\|\mu - \mu_0\|_\infty / \sigma^2 \rightarrow 0$. Under regularity conditions, $\frac{f_0}{f_{\mu,\sigma}}$ has a finite upper bound, hence we can use dominated convergence theorem subsequently. Observe, for a fixed σ , $\int_0^1 \Gamma_\sigma(y - \mu(x)) dx \geq \frac{1}{\sup_{x \in (0,1)} h_\sigma(y, \mu(x) - \mu_0(x))} \int_0^1 \Gamma_\sigma(y - \mu_0(x)) dx$.

As $\|\mu - \mu_0\|_\infty, \sigma^2$ go to 0 with $\|\mu - \mu_0\|_\infty / \sigma^2 \rightarrow 0$, and applying dominated convergence theorem, $0 < \int_{\mathbb{R}} f_0(y) \log \frac{f_0(y)}{f_{\mu,\sigma}(y)} dy \leq \lim_{\sigma \rightarrow 0} \lim_{\|\mu - \mu_0\|_\infty \rightarrow 0} \log(\sup_x h_\sigma(y, \mu(x) - \mu_0(x))) \rightarrow 0$. For Cauchy

errors, $f_0(y) = \lim_{\sigma \rightarrow 0} \int_0^1 \frac{1}{\pi\sigma} \frac{1}{\left(1 + \frac{(y - \mu_0(x))^2}{\sigma^2}\right)} dx$ and

$$f_{\mu,\sigma}(y) = \int_0^1 \frac{1}{\pi\sigma} \frac{1}{\left(1 + \frac{(y - \mu(x))^2}{\sigma^2}\right)} dx = \int_0^1 \frac{1}{\pi\sigma} \frac{1}{\left(1 + \frac{1}{\sigma^2} [(y - \mu_0(x))^2 - 2(y - \mu_0(x))(\mu(x) - \mu_0(x)) + (\mu(x) - \mu_0(x))^2]\right)} dx.$$

As $\|\mu - \mu_0\|_\infty, \sigma^2$ go to 0 with $\|\mu - \mu_0\|_\infty / \sigma^2 \rightarrow 0$, and applying dominated convergence theorem, $\int_{\mathbb{R}} f_0(y) \log \frac{f_0(y)}{f_{\mu,\sigma}(y)} dy \rightarrow 0$. Thus taking appropriate limits, $KL(f_{\mu,\sigma}, f_0)$ goes to 0

under Gaussian, Laplace and Cauchy errors. Hence we can choose a suitably small ϵ_1 and

ϵ_2 with $0 < \epsilon_1 < \epsilon_2$ such that $\left\{ \|\mu - \mu_0\|_\infty \leq \epsilon_1, 0 < \sigma \leq \epsilon_2 \right\} \Rightarrow KL(f_{\mu,\sigma}, f_0) \leq \epsilon$. Using the

assumptions on the support of priors Π^* and ν , we have $\Pi(KL_\epsilon(f_0)) > 0$.

Proof of Theorem 2:

Using the regularity conditions and Taylor's series expansion, we have for fixed y , μ , σ ,

$$\log \frac{f_0(y)}{f_{\mu,\sigma}(y)} = \sum_{k=1}^{n_0} (-1)^k \frac{\{(f_0(y) - 1)^k - (f_{\mu,\sigma}(y) - 1)^k\}}{k} + \delta_1^y(n_0) - \delta_2^y(n_0), \text{ for a fixed } n_0,$$

where $\delta_1^y(n_0) - \delta_2^y(n_0)$ is uniformly bounded in y . Using the identity $a^n - b^n = (a-b)(\sum_{k=1}^n a^{n-k} b^{k-1})$,

and denoting $g_0 = f_0 - 1$ and $g_{\mu,\sigma} = f_{\mu,\sigma} - 1$, we have,

$$\begin{aligned} \int \left| \sum_{k=1}^{n_0} (-1)^k \frac{\{(f_0 - 1)^k - (f_{\mu,\sigma} - 1)^k\}}{k} \right| dy &\leq \sum_{k=1}^{n_0} \int \left| \frac{(-1)^k}{k} (f_0 - f_{\mu,\sigma}) \left(\sum_{l=1}^k g_{\mu,\sigma}^{k-l} g_0^{l-1} \right) \right| dy \\ &\leq \sum_{k=1}^{n_0} \sup_y \left| \frac{(-1)^k}{k} \left(\sum_{l=1}^k g_{\mu,\sigma}^{k-l} g_0^{l-1} \right) \right| \int |f_0(y) - f_{\mu,\sigma}(y)| dy = K(n_0) \int |f_0(y) - f_{\mu,\sigma}(y)| dy, \end{aligned}$$

where $K(n_0) = \sum_{k=1}^{n_0} \sup_y \left| \frac{(-1)^k}{k} \left(\sum_{l=1}^k g_{\mu,\sigma}^{k-l} g_0^{l-1} \right) \right|$ is a finite constant depending on n_0 , for μ belonging to a finite L_1 ball around μ_0 , using the regularity conditions. Further, using similar methods as in the proof of theorem 3, we can show that for $\epsilon_1 < \epsilon_2 < \epsilon_2^*$,

$$\{\mu \in N_{\epsilon_1}(\mu_0), \sigma \in (\epsilon_2, \epsilon_2^*)\} \Rightarrow \int |f_0(y) - f_{\mu,\sigma}(y)| dy < \frac{\epsilon_1}{\epsilon_2}. \quad (8)$$

Using inequality (8), we have for $\mu \in N_{\epsilon_1}(\mu_0)$ and $\sigma \in (\epsilon_2, \epsilon_2^*)$,

$$\begin{aligned} \int \left| \sum_{k=1}^{n_0} (-1)^k \frac{\{(f_0 - 1)^k - (f_{\mu,\sigma} - 1)^k\}}{k} \right| dy &\leq K(n_0) \frac{\epsilon_1}{\epsilon_2}. \text{ Also note, } KL(f_0, f_{\mu,\sigma}) \leq \int f_0 \left| \log \frac{f_0}{f_{\mu,\sigma}} \right| \\ &= \int f_0 \left| \sum_{k=1}^{n_0} (-1)^k \frac{\{(f_0 - 1)^k - (f_{\mu,\sigma} - 1)^k\}}{k} + \delta_1^y(n_0) - \delta_2^y(n_0) \right| \leq K(n_0) \frac{\epsilon_1}{\epsilon_2} + \Delta(n_0) = \epsilon, \end{aligned}$$

for a finite $K(n_0)$ and suitably small $\Delta(n_0) = \sup_y |\delta_1^y(n_0) - \delta_2^y(n_0)|$ with ϵ_1, ϵ_2 depending on ϵ . Under positive L_1 support by Π^* , the rest follows by similar arguments as in theorem 1.

Proof of Theorem 3:

Note that, $|f_{\mu,\sigma}(y) - f_0(y)| \leq |\sup_{x \in (0,1)} \phi_\sigma(y - \mu(x)) - f_0|$, which is integrable $\forall(\mu, \sigma)$.

Then using dominated convergence theorem, $\lim_{\sigma \rightarrow 0} \int |f_{\mu,\sigma} - f_0| = \int \lim_{\sigma \rightarrow 0} |f_{\mu,\sigma} - f_0|$. Now applying Fatou's Lemma and Fubini's Theorem successively, we have,

$$\begin{aligned}
\int \lim_{\sigma \rightarrow 0} |f_{\mu,\sigma} - f_0| dy &\leq \int \lim_{\sigma \rightarrow 0} \int_0^1 |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \mu_0(x))| dx dy \\
&\leq \liminf_{\sigma \rightarrow 0} \int \int_0^1 |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \mu_0(x))| dx dy \quad (\text{Fatou's lemma}) \\
&= \liminf_{\sigma \rightarrow 0} \int_0^1 \int |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \mu_0(x))| dy dx \quad (\text{Fubini's Theorem}) \\
&= \liminf_{\sigma \rightarrow 0} \left\{ \int_{x|\mu_0 > \mu} \int |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \mu_0(x))| dy dx \right. \\
&\quad \left. + \int_{x|\mu_0 < \mu} \int |\phi_\sigma(y - \mu_0(x)) - \phi_\sigma(y - \mu(x))| dy dx \right\}.
\end{aligned}$$

In the proof of lemma 1 of Ghosal, Ghosh and Ramamoorthy (1999), it was shown that for fixed $\theta_1 < \theta_2$, $\|\phi_\sigma(y - \theta_1) - \phi_\sigma(y - \theta_2)\| < \frac{\theta_2 - \theta_1}{\sigma}$, which would imply

$$\begin{aligned}
\int \lim_{\sigma \rightarrow 0} |f_{\mu,\sigma} - f_0| dy &\leq \liminf_{\sigma \rightarrow 0} \left\{ \int_{x|\mu_0 > \mu} \frac{\mu_0(x) - \mu(x)}{\sigma} dx + \int_{x|\mu_0 < \mu} \frac{\mu(x) - \mu_0(x)}{\sigma} dx \right\} \\
&= \liminf_{\sigma \rightarrow 0} \int_0^1 \frac{|\mu(x) - \mu_0(x)|}{\sigma} dx \leq \lim_{\sigma \rightarrow 0} \int_0^1 \frac{|\mu(x) - \mu_0(x)|}{\sigma} dx. \quad (9)
\end{aligned}$$

As $\|\mu - \mu_0\|_\infty, \sigma^2$ go to 0 with $\|\mu - \mu_0\|_\infty / \sigma^2 \rightarrow 0$, the above limit exists and goes to 0. Given that the limit exists and goes to 0, we can now choose sufficiently small $\epsilon_1, \epsilon_2, \epsilon_2^*$ with $0 < \epsilon_1 < \epsilon_2 < \epsilon_2^*$ such that for $\int_0^1 |\mu_0(x) - \mu(x)| dx < \epsilon_1 \equiv \mu \in N_{\epsilon_1}(\mu_0)$ and $\sigma \in (\epsilon_2, \epsilon_2^*)$, we would have $\int |f_{\mu,\sigma} - f_0| dy < \frac{\epsilon_1}{\epsilon_2}$, using (9).

Proof of Theorem 4:

Our proof is based on theorem 2 of Ghosal, Ghosh and Ramamoorthi (1999) who gave a set of alternate sufficient conditions for almost sure convergence of the posterior of strong neighborhoods. Their result involves conditions on the size of the parameter space in terms of L-1 metric entropy. Before proceeding, let us review L-1 metric entropy and theorem 2 of Ghosal, Ghosh and Ramamoorthi (1999).

DEFINITION 1. For $\mathcal{G} \subset \mathcal{F}$ and $\delta > 0$, L-1 metric entropy $J(\delta, \mathcal{G})$ is defined as the minimum of $\log(k : \mathcal{G} \subset \cup_{i=1}^k \{f : \int |f - f_i| dy < \delta, f_1, f_2, \dots, f_k \in \mathcal{F}\})$.

Theorem 5. (Ghosal, Ghosh and Ramamoorthi) Let Π be a prior on \mathcal{F} . Suppose $f_0 \in \mathcal{F}$ is in the Kullback-Leibler support of Π and let $U = \{f : \int |f - f_0| dy < \epsilon\}$. If there is a $\delta < \epsilon/4$, $c_1, c_2 > 0$, $\beta < \epsilon^2/8$ and $\mathcal{F}_n \subset \mathcal{F}$ such that for all large n :

(1) $\Pi(\mathcal{F}_n^c) < c_1 \exp(-nc_2)$, and,

(2) The L-1 metric entropy, $J(\delta, \mathcal{F}_n) < n\beta$,

then $\Pi(U|Y_1, Y_2, \dots, Y_n) \rightarrow 1$ a.s. P_{f_0} .

The constants δ, c_1, c_2, β and \mathcal{F}_n are allowed to depend on ϵ .

Let the parameter space for (μ, σ) be denoted as \mathcal{H} . Consider the subsets of the parameter space $\mathcal{H}_n = \mathcal{H}_{1n} \otimes \mathcal{H}_{2n}$, where $\mathcal{H}_{1n} = \{\mu : \|\mu\|_\infty < M_n, \|\mu'\|_\infty < M_n\}$ and $\mathcal{H}_{2n} = [L_n, \infty)$, with $L_n \rightarrow 0$ such that $\nu(\sigma \in (0, L_n)) < d_1 \exp(-d_2 n)$, $d_1, d_2 > 0$ and $M_n = O(n^{1/2})$. The regularity conditions on the GP guarantees existence of the first derivative μ' with probability 1. Using lemma 4 of Choi and Schervish (2004) who showed an upper bound on sup-norm metric entropy of \mathcal{H}_{1n} , we have the upper bound on L1 metric entropy as

$$J(\delta, \mathcal{H}_{1n}) < K_1 M_n / \delta. \quad (10)$$

This implies there are $K^* = \exp(K_1 M_n / \delta)$ elements $\mu_1, \mu_2, \dots, \mu_{K^*}$ such that

$$\mathcal{H}_{1n} \subset \cup_{j=1}^{K^*} \left\{ \mu : \int_0^1 |\mu - \mu_j| dx < \delta \right\}. \quad (11)$$

Let us consider the sieve $\mathcal{F}_n = \{f_{\mu, \sigma} \in \mathcal{F} : (\mu, \sigma) \in \mathcal{H}_n\}$. Further, let us consider densities $f_{i,n} = f_{\mu_i, L_n} \in \mathcal{F}_n$ defined as in section 2, where the $\mu_i, i = 1, \dots, K^*$ correspond to the ones just defined to cover \mathcal{H}_{1n} . Using similar techniques as in lemma 1 of Ghosal, Ghosh and

Ramamoorthi (1999) and theorem 3, it can be shown that for $f_{\mu,\sigma} \in \mathcal{F}_n$,

$$\int |f_{\mu,\sigma} - f_{i,n}| dy \leq \int_0^1 \frac{|\mu - \mu_i|}{\sqrt{2\pi}\sigma} dx + \int_0^1 \frac{|\mu - \mu_i|}{\sqrt{2\pi}L_n} dx \leq \frac{\sqrt{2}}{\sqrt{\pi}L_n} \int_0^1 |\mu - \mu_i| dx \leq \frac{\delta}{L_n}. \quad (12)$$

This clearly implies

$$J(\delta/L_n, \mathcal{F}_n) = J(\delta, \mathcal{H}_{1n}) < K_1 M_n / \delta \Rightarrow J(\delta, \mathcal{F}_n) \leq K_1 M_n L_n / \delta < n\beta, \quad (13)$$

where we can choose $\delta < \epsilon/4$ such that $\beta < \epsilon^2/8$. Thus the second condition in theorem 6 is satisfied. Also note that the prior probability of \mathcal{F}_n can be calculated in terms of Π^* and ν . Using the regularity conditions on the GP and ν , and similar reasoning as in lemma 5 of Choi and Schervish (2004),

$$\Pi(\mathcal{F}_n^c) = (\Pi^* \otimes \nu)(\mathcal{H}_n^c) \leq c_1 \exp(-c_2 n), c_1, c_2 > 0. \quad (14)$$

Thus the first condition in theorem 6 is satisfied.

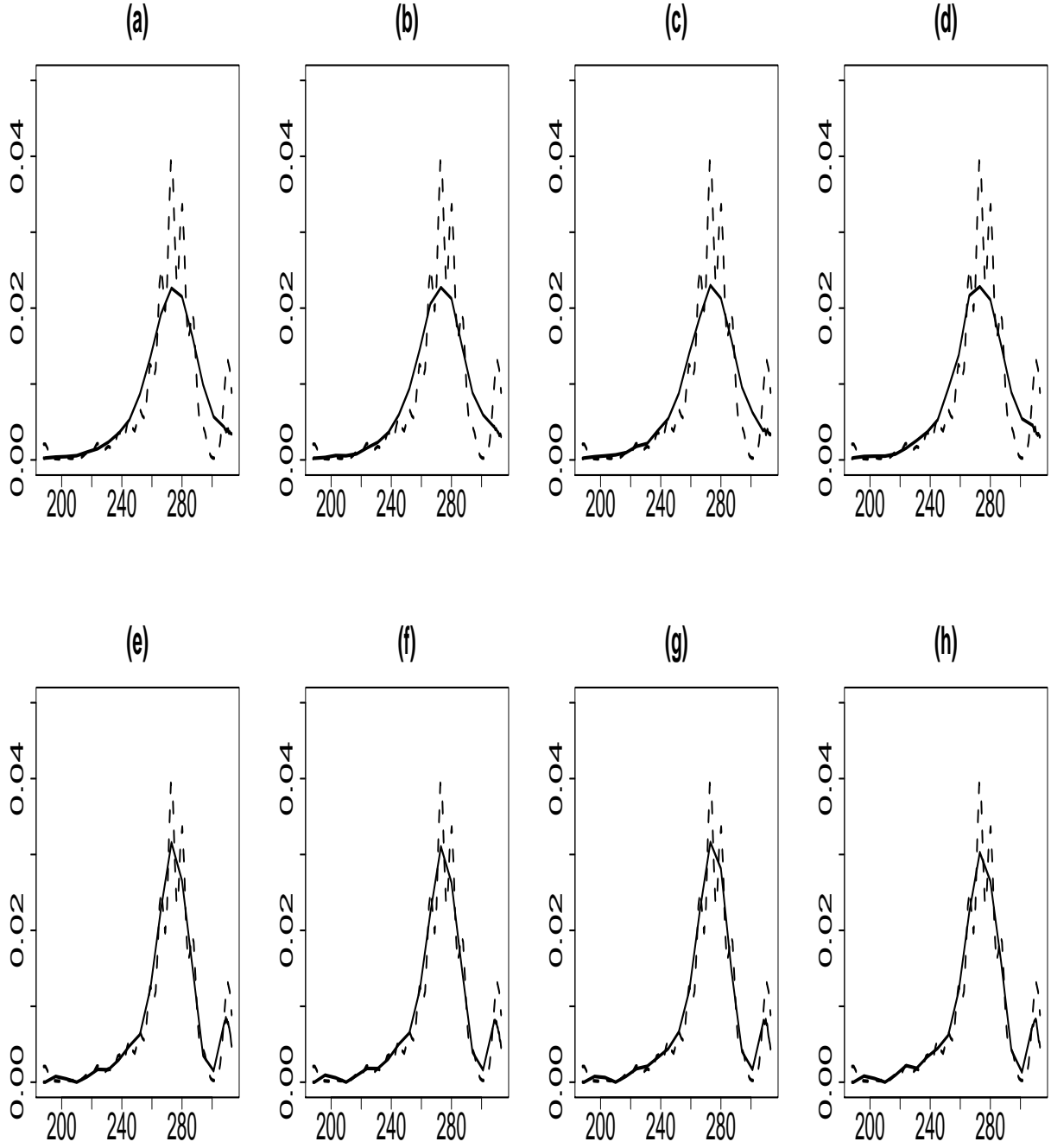


Figure 1: Prior realizations from the GPT for gestational age at delivery (solid lines) along with frequentist kernel density estimate (dotted lines). The rows correspond to $\phi_1=(0.01, 0.1)$; the columns correspond to $\phi_2=(0.1,1,25,100)$.

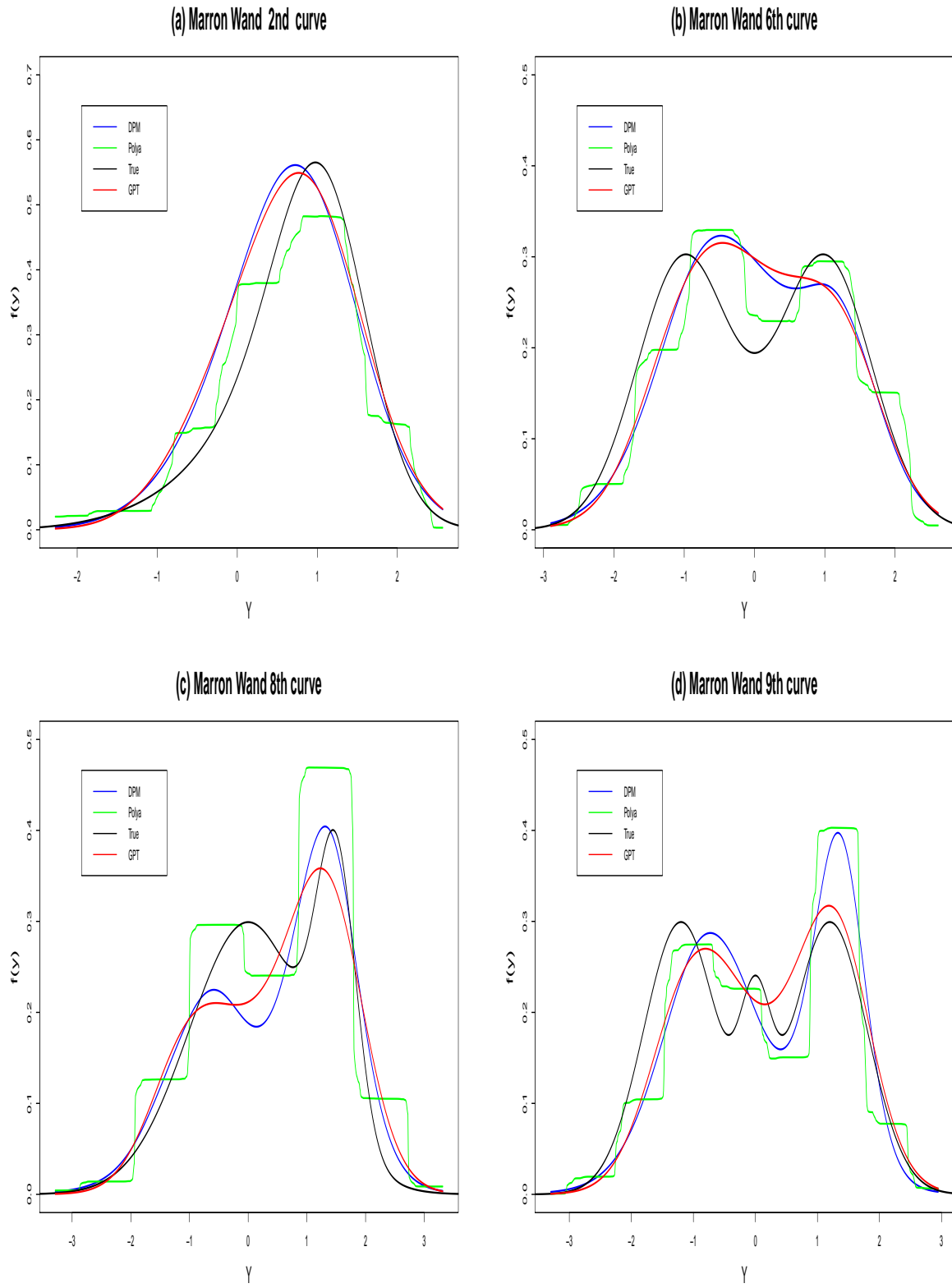


Figure 2: Marron-Wand curves - density estimates for GPT, DPM and Polya tree mixtures.

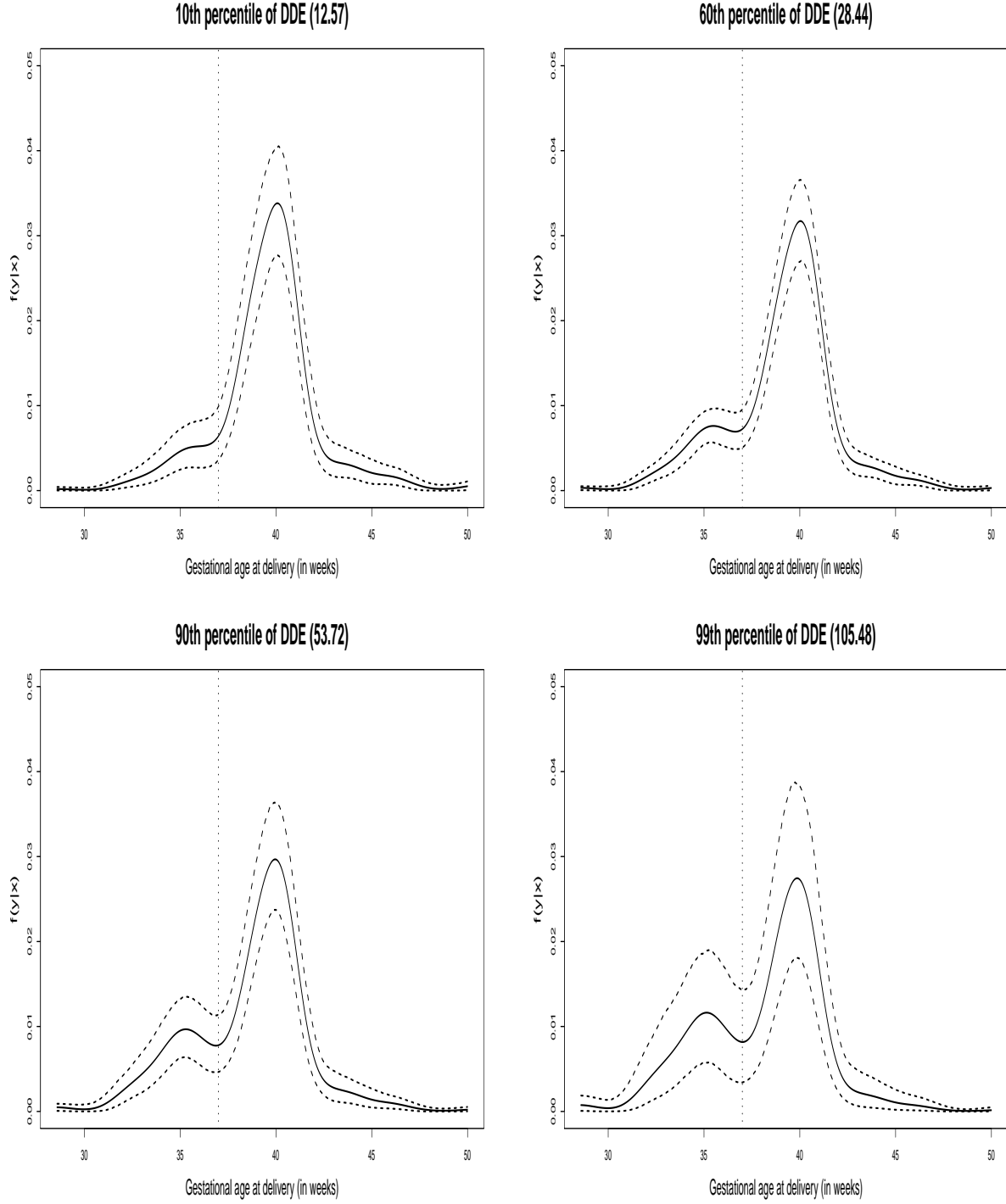


Figure 3: GPT conditional density estimates and 90% credible intervals for 10th, 60th, 90th, 99th DDE quantiles. Vertical dashed line for cut-off at 37 weeks.

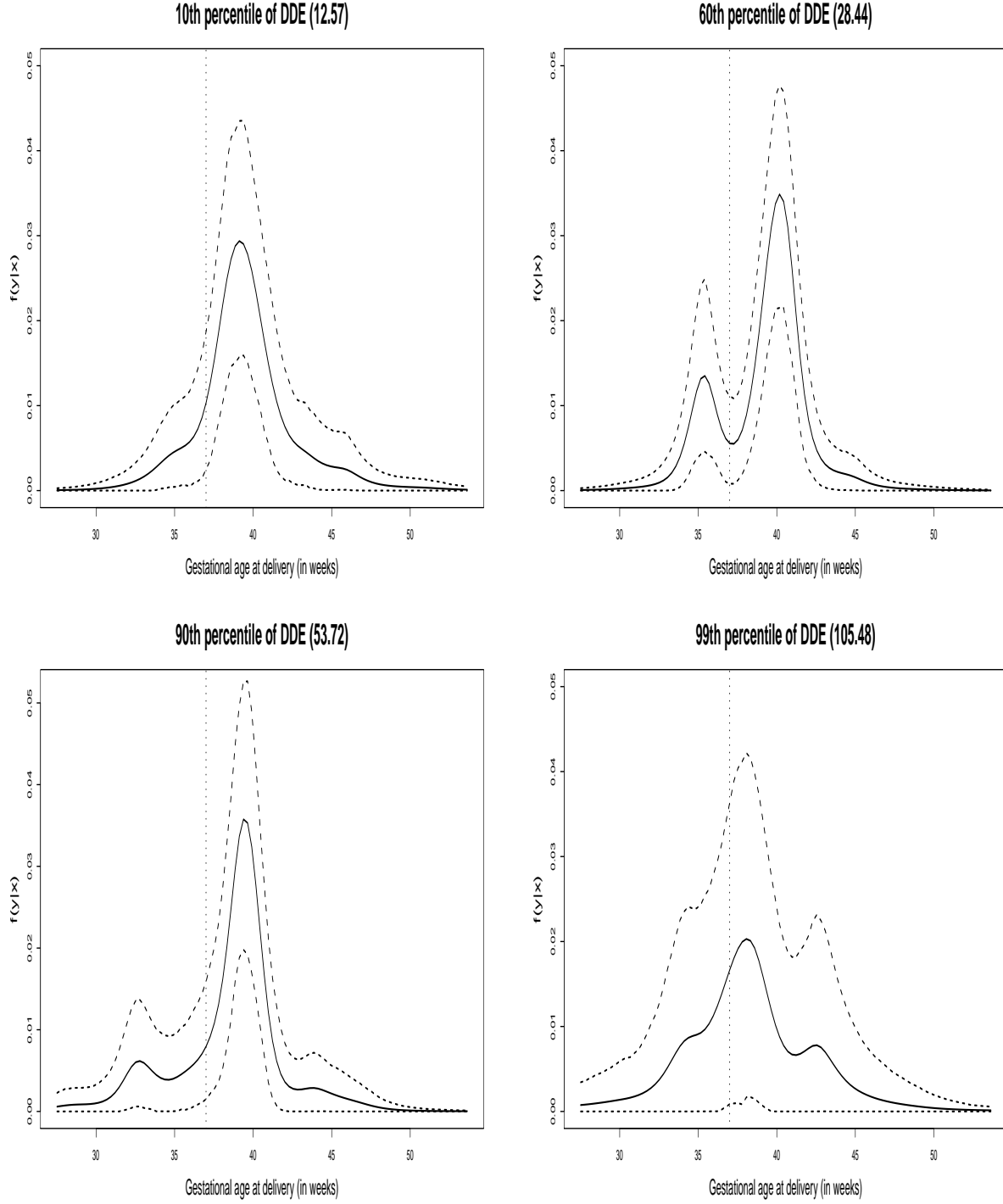


Figure 4: DPM conditional density estimates and 90% credible intervals for 10th, 60th, 90th, 99th DDE quantiles. Vertical dashed line for cut-off at 37 weeks.

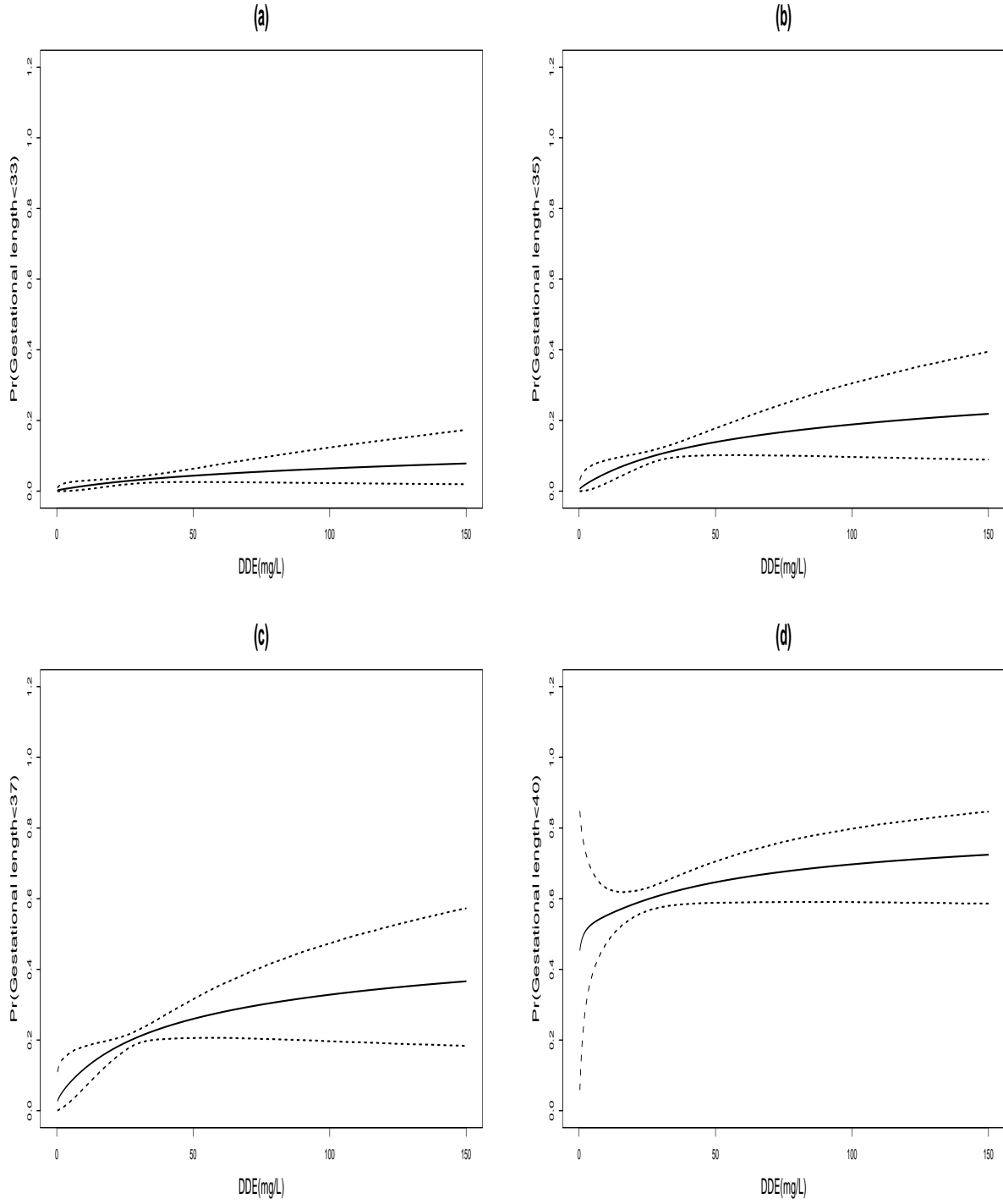


Figure 5: Estimated probability that gestational age at delivery is less than T weeks versus DDE dose, for (a) $T = 33$, (b) $T = 35$, (c) $T = 37$, (d) $T = 40$. Solid lines are posterior means and dashed lines are pointwise 90% credible intervals.

References

- [1] Bush, C.A., and MacEachern, S.N. (1996) A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83, 275 - 285.
- [2] Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q., and West, M. (2008) High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103, 1438 - 1456.
- [3] Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D.B., and Carin, L. (2010) Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: algorithm and performance bounds. *IEEE. Transactions on Signal Processing*, 58, 6140 - 6155.
- [4] Chipman, H., George, E., and McCulloch, R. (2010) BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4, 266 - 298.
- [5] Choi, T., and Schervish, M. (2004) Posterior Consistency in Nonparametric Regression Problems under Gaussian Process Priors. Technical Report.
- [6] De Iorio, M., Muller, P., Rosner, G.L., and MacEachern, S. (2004) An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association*, 99, 205-215.
- [7] Dunson, D.B. (2006) Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, 7, 551 - 568.
- [8] Dunson, D. B., Pillai, N., and Park, J. H. (2007) Bayesian density regression. *Journal of the Royal Statistical Society, Series B*, 69, 163 - 183.
- [9] Dunson, D.B., and Park, J. H. (2008) Kernel stick breaking processes. *Biometrika*, 95, 307 - 323.
- [10] Escobar, M.D., and West, M. (1995) Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90, 577 - 588.
- [11] Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209 - 230
- [12] Ferguson, T. S. (1974) Prior distributions on spaces of probability measures. *Annals of Statistics*, 2, 615 - 629.
- [13] Fokoue, E., and Titterington, D.M. (2003) Mixtures of factor analysers: Bayesian estimation and inference by stochastic simulation. *Machine Learning*, 50, 73 - 94.
- [14] Fokoue, E. (2005) Mixtures of factor analyzers: an extension with covariates. *Journal of Multivariate Analysis*, 95, 370 - 384.
- [15] Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999) Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, 27, 143 - 158.
- [16] Gramacy R.B., and Lee, H. K. H, (2008) Bayesian Treed Gaussian Process Models With an Application to Computer Modeling. *Journal of the American Statistical Association*, 103, 1119 - 1130.
- [17] Griffin, J. E. and Steel, M. F. J. (2006) Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101, 179 - 194.
- [18] Hanson, T. (2006) Inference for Mixtures of Finite Polya Trees. *Journal of the American Statistical Association*, 101, 1548 - 1565.
- [19] Jara, A., and Hanson, T. (2010) A class of mixtures of dependent tail-free processes. *Biometrika*, accepted.

- [20] Kleinman, K.P., and Ibrahim, J.G. (1998) A semiparametric Bayesian approach to the random effects model. *Biometrics*, 54, 921 - 938.
- [21] Lavine, M. (1992) Some Aspects of Polya Tree Distributions for Statistical Modelling. *Annals of Statistics*, 20, 1222 - 1235.
- [22] Lavine, M. (1994) More Aspects of Polya Tree Distributions for Statistical Modelling. *Annals of Statistics*, 22, 1161 - 1176.
- [23] Lawrence, N. (2005) Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *Journal of Machine Learning Research*, 6, 1783 - 1816.
- [24] Lee, S.Y., Lu, B., and Song, X.Y. (2008) Semiparametric Bayesian analysis of structural equation models with fixed covariates. *Statistics in Medicine*, 27, 2341 - 2360.
- [25] Lenk, P. J. (1988) The Logistic Normal Distribution for Bayesian, Nonparametric, Predictive Densities. *Journal of the American Statistical Association*, 83, 509 - 516.
- [26] Lenk, P. J. (1991) Towards a Practicable Bayesian Nonparametric Density Estimator. *Biometrika*, 78, 531 - 543.
- [27] Longnecker, M. P., Klebanoff, M. A., Zhou, H. B., and Brock, J. W. (2001) Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. *Lancet*, 358, 110 - 4.
- [28] MacEachern, S. N. (1999) Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association, 50-55.
- [29] Marron, J. S., and Wand, M. P. (1992) Exact mean integrated squared error. *Annals of Statistics*, 20, 712 - 736.
- [30] Mauldin, R.D., Sudderth, W.D., and Williams, S.C. (1992) Polya Trees and Random Distributions. *Annals of Statistics*, 20, 1203 - 1221.
- [31] Müller, P., Erkanli, A., and West, M. (1996) Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83, 67 - 79.
- [32] Neal, R.M. (2000) Markov Chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9, 249 - 265.
- [33] O'Hagan, A., and Kingman, J. F. C. (1978) Curve Fitting and Optimal Design for Prediction. *Journal of the Royal Statistical Society B*, 40, 1 - 42.
- [34] Ritter, C., and Tanner, M.A. (1992) Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87(419), 861-868.
- [35] Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-50.
- [36] Sheather, S. J., and Jones M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society B*, 53, 683 - 690.
- [37] Silva, R., and Gramacy, R. (2010) Gaussian process structural equation models with latent variables. *Proceedings of the 26th Conference on Uncertainty on Artificial Intelligence, UAI*.
- [38] Schwartz, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete*, 4, 10 - 26.
- [39] Tokdar, S. T., and Ghosh, J. K. (2007) Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137, 34 - 42.
- [40] Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. (2010) Bayesian Density Regression with Logistic Gaussian Process and Subspace Projection Bayesian Analysis. *Bayesian Analysis*, 5, 319 - 344.
- [41] Yang, M., and Dunson, D.B. (2010) Bayesian semiparametric structural equation models with latent variables. *Psychometrika*, 75, 675-693.